Disambiguation using semantic similarity measures

Bibek Behera, David S Batista, Mário J Silva, Francisco M Couto

University of Lisbon, Faculty of Sciences, LaSIGE

Definitions

- Disambiguation – A process to identify the correct reference for a place name from a database. There are two types – multiple places having same name and multiple names having same place-references.
- Semantic similarity measure - The similarity is based on information content. The IC of a concept can be quantified as the negative log likelihood -log p(c) where p(c) is the probability of occurrence of a concept c in a specific corpus

Motivation

- Disambiguation is done by human beings everyday whether it is identifying place names or co-references like “this” or “that”.
- Introduction of a novel technique that uses semantic similarity to remove ambiguities.
- Useful applications like generating geographic signatures for web-documents and developing geographic search engines.

Brief Understanding of the Concept

- Lisboa can be a municipality, district or simply a street.
- Sintra in Lisboa – Lisboa is more likely to be a district
- Human beings disambiguate by knowing the relationships expressed in the context.
- We consider *adjacency* and *part of* relationship

Adjacency or “near” relationship

- Alenquer near Cadaval
- Sintra near Odivelas.
- The relationship adjacency could also be expressed by keywords like north, south, east, west apart from “near”.
- For Example Lisboa is limited to the *north* with Leiria ,to the *east* with Santarém.

“Part of” Relationship

- Santiago is a parish in Sesimbra located in Setubal.
- Alcacer do Sal is a municipality in the distict of setubal.
- Setubal lies in Lisboa,Portugal.
- Part of can be expressed by keywords lies in, within, inside of .

Problem

- The main problem that we see here is disambiguation. Is there a way for computer to implement relationship between places to disambiguate correctly?
- Yes, semantic similarity measure .

Approach

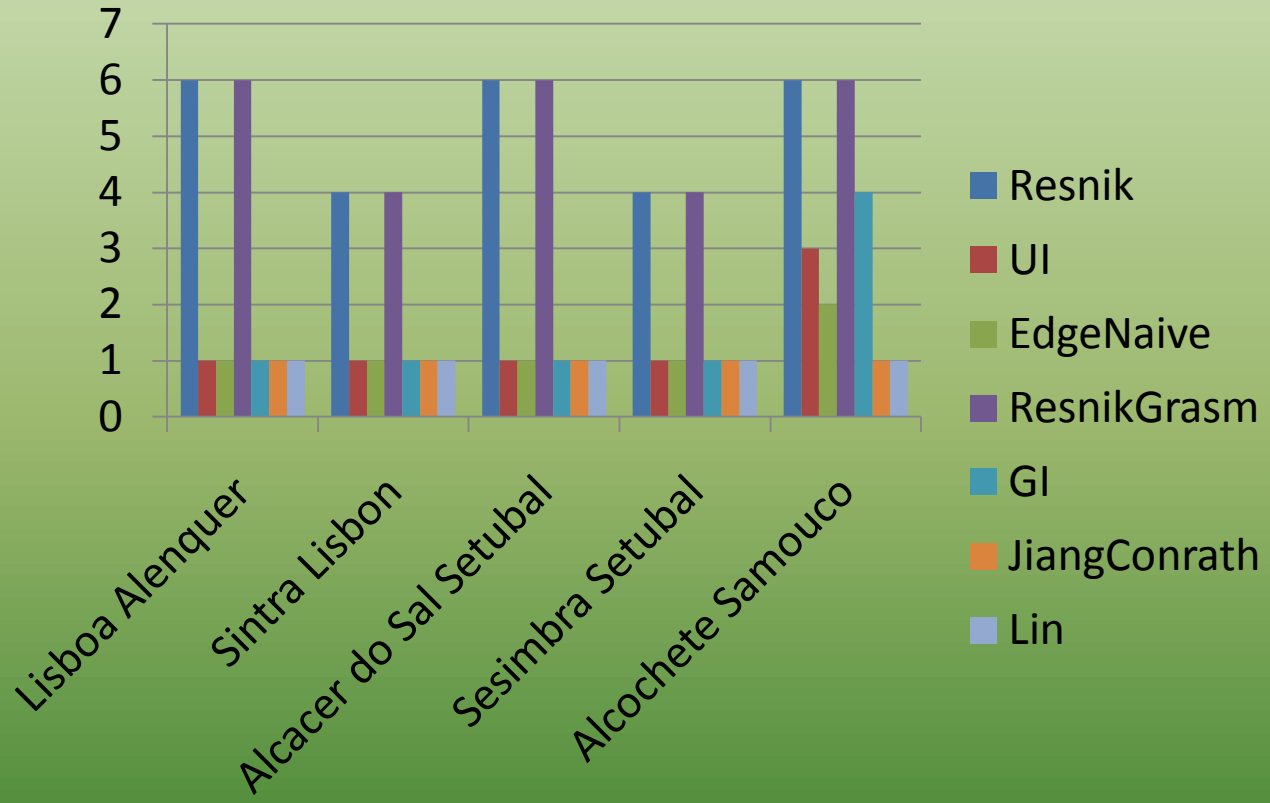
- The notion of maximum similarity .
- For Ex:- Lisbon, Alenquer
- Alenquer has 2 ambiguities
  - Concelho,Locality
- Lisbon has 5 ambiguities
  - ❖ NT2,District,Concelho,Locality ,Locality

*The pairs that have highest semantic similarity are shown for each measure.

- Lin, EdgeNaive
  - DST lisboa-->CON alenquer
- JiangConrath
  - NT2 Lisboa --> CON Alenquer
- Resnik, ResnikGrasm
  - DST lisboa --> CON alenquer
  - DST lisboa --> LOC alenquer
  - CON lisboa --> CON alenquer
  - CON lisboa --> LOC alenquer
  - LOC lisboa --> CON alenquer
  - LOC lisboa --> LOC alenquer

- UI, GI
  - CON Lisboa --> CON Alenquer

Comparision for various measures



Properties of measure

- The UI technique works better for adjacency relationships
- The Lin technique works better for “part of” relationships.
- Resnik and ResnikGrasm perform poorly with lots of noise.
- JiangConrath disambiguates well but incorrect on occasions.
- GI can act as substitute for UI.
- Similarly EdgeNaive can act as substitute for Lin.

Extension of approach for a list of places

- The approach so far described disambiguations for a pair of places.
- Suppose there are a list of places to be disambiguated.
- Disambiguate (Sesimbra, Setubal, Lisboa) using Lin as measure.

Sesimbra	Setubal	Lisboa
Name: Sesimbra type: CON	name: Setubal type: DST	name: Lisboa type: NT2
Name: Sesimbra type: LOC	name: Setubal type: CON	name: Lisboa type: DST
	name: Setubal type: LOC	name: Lisboa type: CON
		name: Lisboa type: LOC
		name: Lisboa type: LOC

Example

- We will consider pairs of place names starting from left.
- Stage 1 result:- Compute similarity between Sesimbra and Setubal
  - CON Sesimbra, DST Setubal
- Stage 2 result:- In these stage we only propagate DST Setubal to compare with place references for Lisboa.
  - DST Setubal , Lisboa NT2
- Finally we merge results from both stages
  - CON Sesimbra --> DST Setubal --> NT2 Lisboa

Preliminary Results

- Precision is 93% using UI as disambiguating tool for “near” relationship.
- Precision is 100% using Lin as disambiguating tool for “part of” relationship.

Conclusion

- Different semantic similarity measures have different effect on disambiguation.
- Using semantic similarity and understanding of relationships, we could disambiguate geo-concepts.

Future Work

- To add a relationship extractor before disambiguation
- To add a ranking algorithm after disambiguation
- Design a query based geographic search engine.
- Generation of geographic signatures.

References

- Marcirio Chaves, Mário J. Silva, Bruno Martins, GKB - Geographic Knowledge Base Technical Report. TR 05-12. FCUL, July 2005.
- A.Joao. Extracting Relations between Concepts and Geo-Temporal Entities in Documents about Art.
- I. Anastacio, B. Martins, and P. Calado. A machine learning method for resolving place references in text. AGILE-09, the 7th European Conference on Geographical Information Science, 2009.
- D.Batista, M.J.Silva, F.M.Couto, B.Behera. Generating Geographic Signatures for Semantic Retrieval. Proceedings of the 6th workshop on GIR, 18-19th Feb. 2010, Zurich, Switzerland.

Thank you

Questions ?